AUDIO-VISUAL SPEECH RECOGNITION WITH A HYBRID SVM-HMM SYSTEM

Mihai Gurban and Jean-Philippe Thiran

Signal Processing Institute, Ecole Polytechnique Fédérale de Lausanne (EPFL)
CH-1015 Lausanne, Switzerland
email: {mihai.gurban, jp.thiran}@epfl.ch
web: itswww.epfl.ch

ABSTRACT

Traditional speech recognition systems use Gaussian mixture models to obtain the likelihoods of individual phonemes, which are then used as state emission probabilities in hidden Markov models representing the words. In hybrid systems, the Gaussian mixtures are replaced by more discriminant classifiers, leading to an improved performance. Most of the time the classifiers used in such systems are neural networks. Support vector machines have also been used in one-modality audio or visual speech recognition, but never in a multimodal audio-visual system. We propose such a hybrid SVM-HMM speech recognizer, and we show how the multimodal approach leads to better performance than that obtained with any of the two modalities individually.

1. INTRODUCTION

In order to understand speech, humans deal with the problem of noise by making use of visual information. It is a well-known fact that human speech perception is bimodal in nature. The visual modality offers important information about the place of articulation, and this information is always used at a subconscious level. The well-known McGurk effect [1] shows that visual stimuli inconsistent with the audio can change the perceived sound.

Audio-visual speech processing promises to take advantage of the same complementarity of the visual and audio modalities, improving recognition rates, especially in the presence of noise, well above those possible with only one modality (an overview of audio-visual speech recognition is given in [2]).

A typical audio-visual speech recognition system represents words with hidden Markov models (HMMs) with each state corresponding to a phoneme. The emission probability of each state is modelled by a mixture of Gaussians, trained with the expectation-maximization algorithm (EM).

But the EM algorithm does not guarantee optimal recognition rates, as it is aimed to model the probability distribution and not provide the best discriminative representation. Replacing the Gaussian mixture with more discriminative classifiers leads to hybrid systems with improved recognition rates [3, 4]. These classifiers could be neural networks or support vector machines (SVMs), although between the two, the SVMs have rarely been used for one-modality systems, and never for multimodal ones. Ganapathiraju [5] reports very good results on audio speech recognition with a hybrid SVM-HMM system, while using only a fraction of the training set used by a traditional system. On the visual part, Gordan [6] obtained a high recognition rate using simple visual features, showing that SVMs are very promising for speech recognition.

To our best knowledge, a hybrid SVM-HMM system has never been used for audio-visual recognition. In our hybrid system, we employ two different audio-visual integration techniques, feature fusion and decision fusion, and we show which of the two works best in the particular case of SVMs.

2. SUPPORT VECTOR MACHINES

According to the empirical risk minimization principle, on which many classifiers are based, the distance between a classifier's outputs and the desired outputs should be minimized for the training examples. While this could produce classifiers that make no errors at all on the training set, the general goal is to classify unseen examples, that is, to generalize. This led to the principle of structural risk minimization, which defines a tradeoff between a classifier's complexity, and the empirical risk [7]. A too high complexity can cause overfitting, a situation where the empirical risk is very small, but the generalization is poor.

Support vector machines are classifiers based on the structural risk minimization principle. They are derived from the optimal hyperplane linear classifiers, maximizing the distance between the separating plane and the closest data points. The larger this distance is, the higher the generalization power of the classifier will be. Since real data is rarely linearly separable, SVMs create a mapping ϕ to a higher dimensional space, the feature space, where a linear separation is sought. The SVM training algorithm requires only inner products in the feature space, which are given by a kernel function, $K(\mathbf{u}, \mathbf{v}) = \langle \phi(\mathbf{u}), \phi(\mathbf{v}) \rangle$. This makes the computation of the mapping ϕ implicit. Typical kernels are polynomials and radial basis functions.

For testing, the distance between any data point and the separating hyperplane can be a good measure of the confidence in the result of classification. In fact, Platt [8] developed a method to transform this distance into a posterior probability, $P(\mathrm{C}|\mathbf{F})$, that is, the probability that the example with feature values $\mathbf{F}$ belongs to class C. His method relies on mapping a sigmoidal function on the outputs of the SVM.

The fact that SVMs use a separating hyperplane makes them binary classifiers. However, groups of SVMs can solve multi-class problems. In our approach, we employ the one-against-one method, building $k(k-1)$ machines, each one discriminating between only two of the k classes. Posterior probabilities are obtained through an optimization performed on the resulting pairwise class probabilities.

But SVMs are also static classifiers, unable to take into account the temporal dependencies which are very important for speech recognition. HMMs can model exactly this dependency. In our case, HMMs are used as word models, while the SVMs serve as phoneme/viseme classifiers. They estimate the probability that one sample belongs to a certain phoneme class.

Our implementation employs the libSVM library [9]. Posterior probabilities are computed with Platt's algorithm [8] mentioned above. The kernels that we use in our experiments are radial basis functions (RBFs).

Figure 1: Two images from the Tulips1 database. The pose differs significantly between the two.

Figure 2: The same two images as in fig. 1, but normalized.

3. THE DATA

The Tulips1 database [10], which was used in our experiments, is a small audio-video database consisting of 12 subjects (9 male, 3 female) saying the first four digits in English. Each digit is pronounced twice by each subject, amounting to a total of 96 words.

The audio part is sampled at 11 kHz with 8 bits per sample. The video part, digitized at 30 frames per second, consists of grayscale images with 8 bits per pixel. The resolution of the images is 100x75. Although only the mouth and the region around it are shown, some movement of the head is allowed (fig. 1). To take this into account, we implement a normalization similar to that reported in [11], using hand-marked points on the lips. The result is that all mouths are now centered, horizontally aligned and scaled to the same horizontal size (fig. 2).

As our system requires the synchronization of the audio and video streams, we limit the audio processing to the speed of the video stream, extracting features 30 times per second. These audio features are extracted from overlapping windows (50% overlap). The size of the window is chosen such that each audio feature vector corresponds to one video frame. The audio features used are 12 mel-cepstral coefficients, together with their first and second derivatives, amounting to a vector of 36 elements. To simulate natural conditions, white Gaussian noise has been added to the audio streams, at different signal to noise ratios (SNRs).

The visual features combine two types of information. The pixels of downsampled images of size 20×15 are coupled with their first temporal derivatives, pixel by pixel differences between consecutive frames. Such features were proven to perform well in conjunction with SVMs [6].

Since labelled data is necessary to train the SVMs, images in half the database have been associated with the corresponding visemes. This has been done through visual inspection, leaving out any ambiguous examples. Audio frames have also been labelled accordingly.

4. OUR HYBRID SYSTEM

4.1 Hidden Markov models

Our system is an isolated word speech recognizer, having left-right HMMs as word models. An example of such a word model is given in fig. 3. Let us define $a_{q_i q_j}$ as the transition probability from state q_i to state q_j. The initial

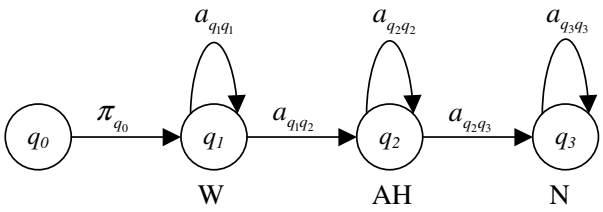

Figure 3: A left-right HMM for the word "one".

transition probability is considered equal to one ($\pi_{q_0} = 1$). The emission probability distribution $b_{q_i}(O_t) = p(O_t|q_i)$ can be obtained from the outputs of the SVMs $p(q_i|O_t)$ through Bayes' rule.

The purpose of the recognition process is to choose the most likely word model, given an observation sequence. The word attached to this model is the recognized word. To this end we employ the Viterbi algorithm [12]. The likelihood of the observation sequence $\mathbf{O} = O_1 O_2 \cdots O_T$, given a path $\mathbf{Q} = q_1 q_2 \cdots q_T$ in the model λ, is:

$$p(\mathbf{O}|\mathbf{Q}, \lambda) = b_{q_1}(O_1) b_{q_2}(O_2) \cdots b_{q_T}(O_T),$$

assuming the statistical independence of the observations. The probability of the path itself is given by:

$$P(\mathbf{Q}|\lambda) = \pi_{q_0} a_{q_1 q_2} a_{q_2 q_3} \cdots a_{q_{T-1} q_T}.$$

The joint probability of $\mathbf{O}$ and $\mathbf{Q}$ occurring simultaneously is the product of the two:

$$p(\mathbf{O}, \mathbf{Q}|\lambda) = \prod_{q_i \in \mathbf{Q}} b_{q_i}(O_i) \cdot \pi_{q_0} \prod_{(q_i, q_j) \in \mathbf{Q}} a_{q_i q_j}.$$

The likelihood of the observation sequence given the model is the sum of these joint probabilities over all possible state sequences $\mathbf{Q}$ [3]:

$$p(\mathbf{O}|\lambda) = \sum_{\text{all } \mathbf{Q}} p(\mathbf{O}, \mathbf{Q}|\lambda)).$$

This "full" likelihood can be replaced by the "Viterbi" approximation [3], considering only the most likely path in the model:

$$p(\mathbf{O}|\lambda) \simeq \max_{\mathbf{Q}} [p(\mathbf{O}, \mathbf{Q}|\lambda)].$$

This often-used approximation does not lead to a significant performance loss, while facilitating the numerical computation. As a further simplification, we can consider the state transition probabilities $a_{q_i q_j}$ as equal and ignore them altogether. This makes sense as, numerically, the emission likelihoods have a much larger influence on the result. Since $\pi_{q_0} = 1$, the joint probability of $\mathbf{O}$ and $\mathbf{Q}$ becomes:

$$p(\mathbf{O}, \mathbf{Q}|\lambda) = \prod_{q_i \in \mathbf{Q}} b_{q_i}(O_i).$$

In the end, the recognized word is given by the most likely word model:

$$\lambda_{\text{recognized}} = \arg \max_{\lambda} [p(\mathbf{O}|\lambda)].$$

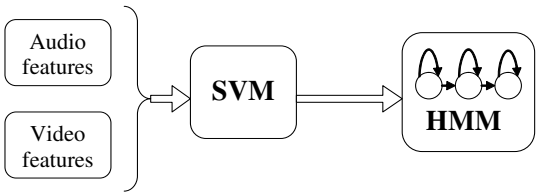

Figure 4: A hybrid SVM-HMM system with feature fusion.

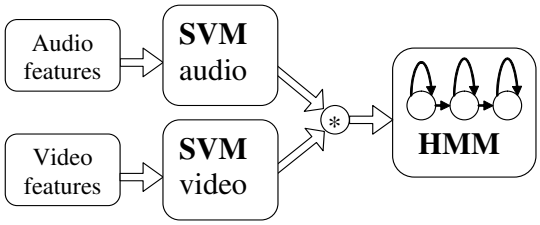

Figure 5: A hybrid SVM-HMM system with decision fusion.

4.2 Audio-visual integration

Several different methods can be used for the integration of audio and visual information [2]. The simplest one is feature concatenation [13], where the audio and video feature vectors are simply concatenated before being presented to the SVM classifier (fig. 4). Here, a single classifier is trained with combined data from the two modalities.

A more complex way of integration is decision fusion, using two classifiers, one for each modality [2]. Our approach is early integration, at frame level, having two different SVM classifiers, whose outputs will be combined before being presented to the single HMM. The posterior probabilities $P(\mathrm{C}|\mathbf{F}_A)$ and $P(\mathrm{C}|\mathbf{F}_V)$ are weighted with a factor α and then multiplied:

$$P(\mathrm{C}|\mathbf{F}_{AV}) = P(\mathrm{C}|\mathbf{F}_A)^{\alpha} \cdot P(\mathrm{C}|\mathbf{F}_V)^{1-\alpha}$$

This expression can be interpreted as a weighted average in the logarithmic domain. The product rule is one of the most widely used probability combination rules, along with the sum rule, the min rule or the max rule [14]. These rules are compared in [15], with the purpose of combining the outputs of classifiers trained on different types of audio-only features. The product rule was found to be the best performer. The same weighted product rule can be found in [13], integrating word-level probabilities.

5. EXPERIMENTAL RESULTS

All our experiments were done using the leave-one-out testing procedure. Training was done on eleven of the twelve speakers in the data set, while testing was performed on the last one. The procedure was repeated for each one of the twelve speakers, and the results were averaged.

As noise was added only in the audio, the video-only performance was constant with respect to the SNR. The two different types of audio-visual integration were both tested for all audio SNRs.

Simple feature fusion was disappointing, in the sense that, for clean audio, the multimodal system performed worse than the audio-only one (fig. 6). However, for lower audio SNRs, the multimodal word recognition rate was higher than the ones obtained with either the audio or the video.

For decision fusion, the weighting factor α has been tuned manually for each SNR. The results were very good for high

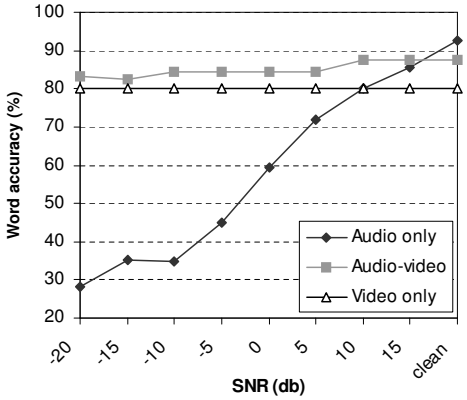

Figure 6: Word accuracy with feature fusion.

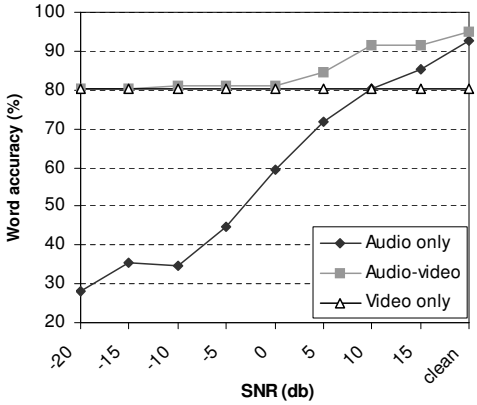

Figure 7: Word accuracy with decision fusion.

SNRs (fig. 7), achieving big gains from multimodal processing. The best example is the 10 dB level, where both audio and video had a performance of around 80% individually, while when put together this grew to 91%. This certainly proves the validity of the multimodal approach.

Unfortunately, in the case of low SNRs, the same behavior could not be replicated. The performance dropped to the level of the video-only system.

The three graphs in fig. 8 represent the variation of the word recognition rate for different weighting factors α. The predictable result is that the weight that should be assigned to the audio is linked to its SNR. A higher SNR means more importance should be given to the audio stream.

There is a point on each of the three graphs where the two modalities seem to be complementary, and the combined result is better than for each of them individually. An automatic method could be derived to find the weight α corresponding to this point. We could rely on confidence values for the one-modality SVMs, in particular the dispersions of the phoneme/viseme likelihoods. Such dispersions have been used at word level in [13]. Unfortunately, our experiments with this method show that dispersion is here a poor measure of confidence. The explanation could be that we assigned more viseme classes to the same sound, as the same viseme can differ significantly from speaker to speaker [6].

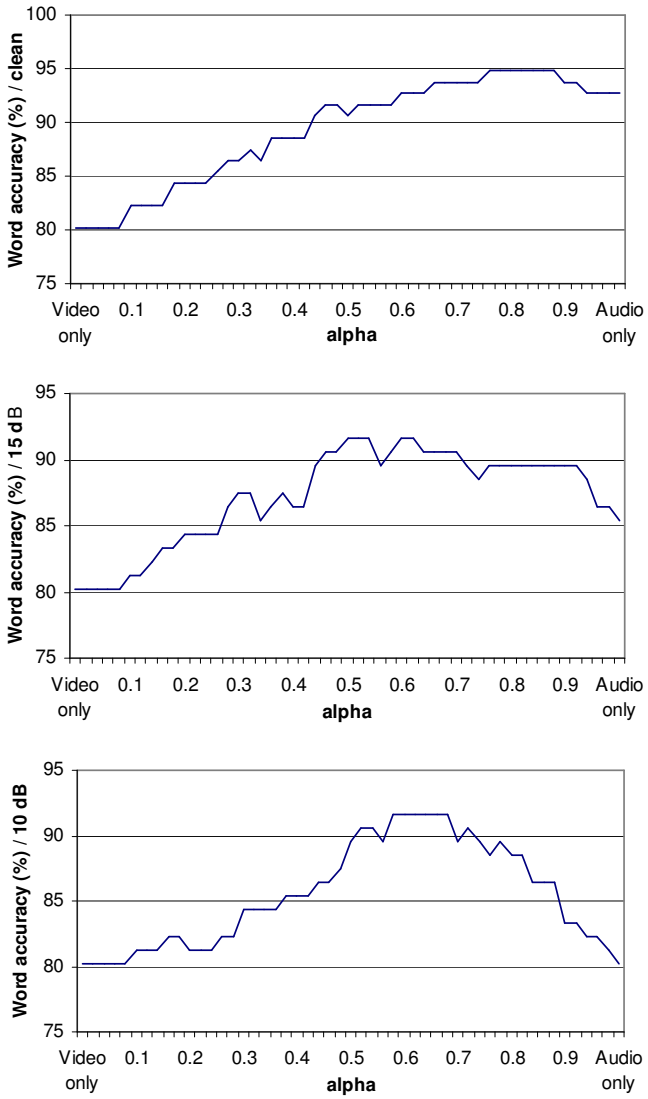

Figure 8: The evolution of word accuracy with the weighting factor α for three different SNR levels.

This lowers the dispersion of the viseme likelihoods, decreasing the confidence. Possible replacements for this confidence measure could be the audio SNR or amount of voicing, both estimated from the audio stream [2].

6. CONCLUSION

Hybrid speech recognition systems perform better than simple HMM-based ones. In this context, although rarely used, SVMs prove to be a real alternative to neural networks. We have successfully integrated SVMs and HMMs into a system that is able to process multimodal data. For this, we used multi-class SVM classifiers with probabilistic outputs, one classifier in the case of feature fusion, and two for decision fusion.

We obtained good accuracy even with the simplest visual features, showing the discriminating power of SVMs. Most of the time, the combination of the two modalities lead to better performance than any of them individually, proving that audio and visual information are complementary in speech.

The way in which the modalities are combined is also relevant. We showed that the ability to assign different weights to each modality, accounting for different environmental conditions, is an important factor in the design of audio-visual speech recognizers.

Acknowledgement

This work is supported by the Swiss National Science Foundation through the IM2 NCCR.

REFERENCES

[1] H. McGurk and J. MacDonald, "Hearing lips and seeing voices," *Nature*, vol. 264, pp. 746–748, 1976.

[2] G. Potamianos, C. Neti, J. Luettin, and I. Matthews, "Audio-visual automatic speech recognition: an overview," in *Issues in audio-visual speech processing* (G. Bailly, E. Vatikiotis-Bateson, and P. Perrier, eds.), MIT Press, 2004.

[3] N. Morgan and H. Bourlard, "Continuous speech recognition, an introduction to the hybrid HMM/connectionist approach," *IEEE Signal Processing Magazine*, vol. 12, no. 3, pp. 25–42, 1995.

[4] H. Bourlard and N. Morgan, *Connectionist Speech Recognition - A Hybrid Approach*. Kluwer Academic Publishers, 1994.

[5] A. Ganapathiraju, J. Hamaker, and J. Picone, "Hybrid SVM/HMM architectures for speech recognition," *International Conference on Spoken Language Processing*, vol. 4, pp. 504–507, 2000.

[6] M. Gordan, C. Kotropoulos, and I. Pitas, "A support vector machine-based dynamic network for visual speech recognition applications," *EURASIP Journal on Applied Signal Processing*, vol. 2002(11), pp. 1248–1259, 2002.

[7] V. Vapnik, *The nature of statistical learning theory*. Springer, 1995.

[8] J. Platt, "Probabilistic outputs for support vector machines and comparison to regularized likelihood methods," in *Advances in Large Margin Classifiers* (A. Smola, P. Bartlett, B. Schoelkopf, and D. Schuurmans, eds.), pp. 61–74, MIT Press, 2000.

[9] C.-C. Chang and C.-J. Lin, *LIBSVM: a library for support vector machines*, 2001. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[10] J. R. Movellan, "Visual speech recognition with stochastic networks," in *Advances in Neural Information Processing Systems* (G. Tesauro, D. Touretzky, and T. Leen, eds.), vol. 7, MIT Press, 1995.

[11] J. Luettin and N. A. Thacker, "Speechreading using probabilistic models," *Computer Vision and Image Understanding*, vol. 65(2), pp. 163–178, 1997.

[12] L. R. Rabiner, "A tutorial on Hidden Markov Models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77(2), 1989.

[13] A. Adjoudani and C. Benoît, "On the integration of auditory and visual parameters in an HMM-based ASR," in *Speechreading by humans and machines* (D. G. Stork and M. E. Hennecke, eds.), pp. 461–471, Springer, 1996.

[14] J. Kittler, M. Hatef, R. Duin, and J. Matas, "On combining classifiers," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 3, pp. 226–239, 1998.

[15] K. Kirchhoff and J. Bilmes, "Dynamic classifier combination in hybrid speech recognition systems using utterance-level confidence values," *Proceedings ICASSP-99*, pp. 693–696, 1999.